CLAIMS

SUB A17

1. A computer processing apparatus for classifying a document, comprising:

means for accessing a database structure providing a plurality of different subject matter categories, the database containing a classified vocabulary consisting of terms in all of the different subject matter categories with each term being classified in accordance with the subject matter category structure of the database;

means for receiving in computer-readable form a text document to be classified;

processor means operable to compare terms appearing in the text document with the terms in the classified vocabulary and to determine from the comparison the category for the document; and

means for supplying a signal carrying data representing the text document and data associating the text document with the determined category.

2. A computer processing apparatus for checking spelling in a document, comprising:

means for accessing a database structure providing a plurality of different subject matter categories, the database containing a classified vocabulary consisting of

terms in all of the different subject matter categories with each term being classified in accordance with the subject matter category structure of the database;

means for receiving in computer-readable form a text document to be spell-checked;

processor means operable to compare terms appearing in the text document with the terms in the classified vocabulary, to determine from the comparison the category for the document, to identify any term in the document not present in the classified vocabulary and to determine the term or terms in the classified vocabulary closest to an unidentified term and having the same category as that determined for the document; and

means for supplying a user with said determined term or terms.

3. A computer processing apparatus for refining the results of a subject matter search carried out by a search engine using a keyword, the apparatus comprising:

means for accessing a database having a database structure providing a plurality of different subject matter categories, the database containing a classified vocabulary consisting of terms in all of the different subject matter categories with each term being classified in accordance with the subject matter category structure

of the database;

means for receiving in computer-readable form documents forming the results of the subject matter search;

processor means operable to compare the keyword used to carry out the search with the classified vocabulary to determine each category with which the keyword is associated;

means for advising a user of the different categories with which the keyword is associated;

user-operable means selection means for enabling a user to select one of said different categories;

means for comparing the terms used in the search result documents with the terms in the classified vocabulary; and

means for supplying the user with information relating the search results to the selected category.

SUB A27

4. Apparatus according to claim 1, wherein the processor means is operable to determine the category for the document by determining from the comparison the category or categories of terms in the document, assigning weightings to the determined categories for the terms, and assigning the document being classified to the category having the highest weighting.

5. Apparatus according to claim 2, wherein the processor means is operable to determine the category for the document by determining from the comparison the category or categories of terms in the document, assigning weightings to the determined categories for the terms, and assigning the document being classified to the category having the highest weighting.

6. Apparatus according to claim 3, wherein the processor means is operable to determine the category for the document by determining from the comparison the category or categories of terms in the document, assigning weightings to the determined categories for the terms, and assigning the document being classified to the category having the highest weighting.

SUB A37

7. Apparatus according to claim 4, wherein the processor means is operable, for each term in the classified vocabulary and in the text document, to share a predetermined weighting factor between each category associated with the term.

8. Apparatus according to claim 1, wherein the accessing means is arranged to access a plurality of collocations also forming part of the database, each

collocation being associated with a specific different one of the subject matter categories and each collocation consisting of a plurality of terms exemplifying the associated category.

9. Apparatus for generating a database for storage on a computer-readable medium, comprising:

means for storing terms;

means for associating each term with one of a number of different subject matter categories;

means for associating all terms falling within the same category with a common code identifying a collocation of terms exemplifying that category so that terms in different categories are associated with different codes identifying different collocations with each collocation being specific to the associated category; and

means for supplying as a database each term together with the associated code.

10. Apparatus according to claim 9, further comprising means storing said collocations.

11. Apparatus according to claim 9, wherein the supplying means is arranged also to supply the

collocations with the database.

SUB A4 12. A computer processing apparatus for classifying a document, comprising:

means for accessing a database having a database structure providing a plurality of different subject matter categories, the database containing a classified vocabulary consisting of terms in all of the different subject matter categories with each term being classified in accordance with the subject matter category structure of the database and the database also containing a plurality of collocations each collocation being associated with a specific different one of the subject matter categories and each collocation consisting of a plurality of terms exemplifying the associated category;

means for receiving in computer-readable form a text document to be classified;

processor means operable to compare terms appearing in the text document with the collocations to determine the collocation having the most terms in common with the document, and to allocate the category of the determined collocation to the document; and

means for supplying a signal carrying data representing the text document and data associating the text document with the determined category.

13. A computer processing apparatus for checking spelling in a document, comprising:

means for accessing a database having a database structure providing a plurality of different subject matter categories, the database containing a classified vocabulary consisting of terms in all of the different subject matter categories with each term being classified in accordance with the subject matter category structure of the database and the database also containing a plurality of collocations each collocation being associated with a specific different one of the subject matter categories and each collocation consisting of a plurality of terms exemplifying the associated category;

means for receiving in computer-readable form a text document to be spell-checked;

processor means operable to compare terms appearing in the text document with the collocations to determine the collocation having most terms in common with the text document, to select the category of that collocation as the category for the document, to identify any term in the document not present in the classified vocabulary and to determine the term or terms in the classified vocabulary closest to an unidentified term and having the same category as that determined for the document; and

means for advising a user of the determined term or

terms.

14 A computer processing apparatus for refining the results of a subject matter search carried out by a search engine using a keyword, the apparatus comprising:

means for accessing a database having a database structure providing a plurality of different subject matter categories, the database containing a classified vocabulary consisting of terms in all of the different subject matter categories with each term being classified in accordance with the subject matter category structure of the database and the database also containing a plurality of collocations each collocation being associated with a specific different one of the subject matter categories and each collocation consisting of a plurality of terms exemplifying the associated category;

means for receiving in computer-readable form documents forming the results of the subject matter search;

processor means operable to compare the keyword used to carry out the search with the classified vocabulary to determine each category with which the keyword is associated;

means for advising a user of the different categories with which the keyword is associated;

user-operable means selection means for enabling a user to select one of said different categories;

means for accessing the collocation associated with the selected category;

means for comparing the terms used in the search result documents with the terms in the accessed collocation; and

means for supplying the user with information relating the search results to the selected category.

15. Apparatus according to claim 14, wherein the supplying means is arranged to supply the user with details of the search results having greater than a predetermined number of terms in common with the accessed collocation.

16. Apparatus according to claim 9, wherein the processor means is operable to disambiguate between different meanings of terms by using the collocations.

SUB B5 17. Apparatus according to claim 12, wherein the processor means is operable to disambiguate between different meanings of terms by using the collocations.

18. Apparatus according to claim 14, wherein the

processor means is operable to disambiguate between different meanings of terms by using the collocations.

SUB A5 19. Apparatus according to claim 7, wherein the accessing means is arranged to access the collocations from store means separate from the remainder of the database.

20. Apparatus according to claim 1, further comprising store means configured to store the database.

21. Apparatus according to claim 1, further comprising store means storing the database.

22. Apparatus according to claim 1, wherein the database structure provides said plurality of subject matter categories as a tree structure consisting of a plurality of main subject matter areas each divided into two or more subsidiary subject matter areas.

23. Apparatus according to claim 1, wherein the database structure provides said plurality of subject matter categories such that each category is defined by a subject matter area and a species or genus.

24. Apparatus according to claim 23, wherein the database provides said plurality of subject matter categories such that the species or geni are people, places, organisations, products and technology.

25. Apparatus according to claim 23, wherein the database structure provides said plurality of subject matter categories such that the species or genus are the same for each subject matter area.

26. Apparatus according to claim 1, wherein the database provides categories in each of the following subject matter areas: the universe, the earth, the environment, natural history, humanity, recreation, society, the mind and human history.

27. Apparatus according to claim 1, wherein the database structure is such that, for a given meaning, a term is associated with only one category and different meanings of the same term are associated with different categories.

28. Apparatus according to claim 1, wherein the supplying means comprises means for storing a signal supplied by the supplying means on a computer readable

medium.

29. Apparatus according to claim 1, wherein the supplying means comprises means for forwarding a signal supplied by the supplying means to another processing apparatus.

30. Apparatus according to claim 1, wherein the supplying means comprises means for displaying the information to a user.

31. In a computer processing apparatus having means for accessing a database having a database structure providing a plurality of different subject matter categories, the database containing a classified vocabulary consisting of terms in all of the different subject matter categories with each term being classified in accordance with the subject matter category structure of the database and means for receiving in computer-readable form a text document to be classified, a method of classifying documents comprising:

comparing terms appearing in the text document with the terms in the classified vocabulary;

determining from the comparison the category for the document; and

supplying a signal carrying data representing the text document and data associating the text document with the determined category.

32. In a computer processing apparatus having means for accessing a database having a database structure providing a plurality of different subject matter categories, the database containing a classified vocabulary consisting of terms in all of the different subject matter categories with each term being classified in accordance with the subject matter category structure of the database and means for receiving in computer-readable form a text document to be spell-checked, a method of checking spelling in a document comprising:

comparing terms appearing in the text document with the terms in the classified vocabulary;

determining from the comparison the category for the document;

identifying any term in the document not present in the classified vocabulary;

determining the term or terms in the classified vocabulary closest to an unidentified term and having the same category as the document; and

advising a user of the determined term or terms.

33. In computer processing apparatus having means for accessing a database having a database structure providing a plurality of different subject matter categories, the database containing a classified vocabulary consisting of terms in all of the different subject matter categories with each term being classified in accordance with the subject matter category structure of the database and means for receiving in computer-readable form documents forming the results of the subject matter search, a method of refining the results of a subject matter search carried out by a search engine using a keyword, the method comprising:

comparing the keyword used to carry out the search with the classified vocabulary to determine each category which the keyword is associated;

advising a user of the different categories with which the keyword is associated;

identifying the one of said categories selected by a user using user-operable selection means;

comparing the terms used in the search result documents with the terms in the classified vocabulary; and

supplying the user with information relating the search results to the selected category.

SUB A47

34. A method according to claim 31, comprising determining the category for the document by determining from the comparison the category or categories of the terms in the document, assigning weightings to the determined categories for the terms, and assigning the document being classified to the category having the highest weighting.

35. A method according to claim 34, which comprises assigning weightings by, for each term in the classified vocabulary and in the text document, sharing a predetermined weighting factor between each category associated with the term.

36. A method according to claim 31 which also comprises accessing a plurality of collocations also forming part of the database, each collocation being associated with a specific different one of the subject matter categories and each collocation consisting of a plurality of terms exemplifying the associated category.

37. In a computer apparatus having data storage means, a method of generating a database for storage on a computer readable medium, comprising:

storing terms;

associating each term with one of a number of different subject matter categories;

associating all terms falling within the same category with a common code identifying a collocation of terms exemplifying that category so that terms in different categories are associated with different codes identifying different collocations with each collocation being specific to the associated category; and

supplying as a database each term together with the associated code.

38. A method according to claim 37, which comprises supplying the collocations with the database.

SUB A7

39. In a computer processing apparatus having means for accessing a database having a database structure providing a plurality of different subject matter categories, the database containing a classified vocabulary consisting of terms in all of the different subject matter categories with each term being classified in accordance with the subject matter category structure of the database and the database also containing a plurality of collocations each collocation being associated with a specific different one of the subject matter categories and each collocation consisting of a

plurality of terms exemplifying the associated category and having means for receiving in computer-readable form a text document to be classified, a method of classifying documents comprising:

comparing terms appearing in the text document with the collocations to determine the collocation having the most terms in common with the document;

allocating the category of the determined collocation to the document; and

supplying a signal carrying data representing the text document and data associating the text document with the determined category.

40. In a computer processing apparatus having means for accessing a database having a database structure providing a plurality of different subject matter categories, the database containing a classified vocabulary consisting of terms in all of the different subject matter category structure of the database and the database also containing a plurality of collocations each collocation being associated with a specific different one of the subject matter categories and each collocation consisting of a plurality of terms exemplifying the associated category and having means for receiving in computer-readable form a text document to be spell-

checked, a method of checking spelling in a document comprising:

comparing terms appearing in the text document with the collocations to determine the collocation having most terms in common with the text document;

selecting the category of that collocation as the category for the document;

identifying any term in the document not present in the classified vocabulary;

determining the term or terms in the classified vocabulary closest to an unidentified term and having the same category as that selected for the document; and

advising a user of the determined term or terms.

41. In a computer processing apparatus having means for accessing a database having a database structure providing a plurality of different subject matter categories, the database containing a classified vocabulary consisting of terms in all of the different subject matter categories with each term being classified in accordance with the subject matter category structure of the database and the database also containing a plurality of collocations each collocation being associated with a specific different one of the subject matter categories and each collocation consisting of a

plurality of terms exemplifying the associated category, and having means for receiving in computer-readable form documents forming the results of a subject matter search carried out by a search engine using a keyword, a method of refining the search results, comprising:

comparing the keyword used to carry out the search with the classified vocabulary to determine each category with which the keyword is associated;

advising a user of the different categories with which the keyword is associated;

determining which of said categories is selected by a user using user-operable means selection means;

accessing the collocation associated with the selected category;

comparing the terms used in the search result documents with the terms in the accessed collocation; and

supplying the user with information relating the search results to the selected category.

42. A method according to claim 41, which comprises supplying the user with details of the search results having greater than a predetermined number of terms in common with the accessed collocation.

SUB A8 43. A method according to claim 36, which comprises

accessing the collocations from store means separate from the remainder of the database.

44. A method according to claim 37, which comprises structuring the database to provide said plurality of subject matter categories as a tree structure consisting of a plurality of main subject matter areas each divided into two or more subsidiary subject matter areas.

45. A method according to claim 37, which comprises structuring the database to provide said plurality of subject matter categories such that each category is defined by a subject matter area and a species or genus.

46. A method according to claim 45, which comprises structuring the database to provide said plurality of subject matter categories such that the species or geni are people, places, organisations, products and technology.

47. A method according to claim 45, which comprises structuring the database structure to provide said plurality of subject matter categories such that the species or genus are the same for each subject matter area.

48. A method according to claim 37, which comprises structuring the database to provide categories in each of the following subject matter areas: the universe, the earth, the environment, natural history, humanity, recreation, society, the mind and human history.

49. A method according to claim 37, which comprises structuring the database such that, for a given meaning, a term is associated with only one category and different meanings of the same term are associated with different categories.

SUB A97

50. A method according to claim 31, which comprises carrying out the supplying by storing a signal on a computer-readable medium.

51. A method according to claim 31, which comprises carrying out the supplying by forwarding a signal to another processing apparatus.

52. A method according to claim 31, which comprises displaying the information to a user.

53. A database for use with an apparatus in accordance with claim 1, the database having a database structure

providing a plurality of different subject matter categories, the database containing a classified vocabulary consisting of terms in all of the different subject matter categories with each term being classified in accordance with the subject matter category structure of the database.

54. A database for use with an apparatus in accordance with claim 2, the database having a database structure providing a plurality of different subject matter categories, the database containing a classified vocabulary consisting of terms in all of the different subject matter categories with each term being classified in accordance with the subject matter category structure of the database.

55. A database for use with an apparatus in accordance with claim 3, the database having a database structure providing a plurality of different subject matter categories, the database containing a classified vocabulary consisting of terms in all of the different subject matter categories with each term being classified in accordance with the subject matter category structure of the database.

SUB A10

56. A database for use with an apparatus in accordance with claim 12, the database having a database structure providing a plurality of different subject matter categories, the database containing a classified vocabulary consisting of terms in all of the different subject matter categories with each term being classified in accordance with the subject matter category structure of the database and the database also containing a plurality of collocations each collocation being associated with a specific different one of the subject matter categories and each collocation consisting of a plurality of terms exemplifying the associated category.

57. A database for use with an apparatus in accordance with claim 13, the database having a database structure providing a plurality of different subject matter categories, the database containing a classified vocabulary consisting of terms in all of the different subject matter categories with each term being classified in accordance with the subject matter category structure of the database and the database also containing a plurality of collocations each collocation being associated with a specific different one of the subject matter categories and each collocation consisting of a plurality of terms exemplifying the associated category.

58. A database for use with an apparatus in accordance with claim 14, the database having a database structure providing a plurality of different subject matter categories, the database containing a classified vocabulary consisting of terms in all of the different subject matter categories with each term being classified in accordance with the subject matter category structure of the database and the database also containing a plurality of collocations each collocation being associated with a specific different one of the subject matter categories and each collocation consisting of a plurality of terms exemplifying the associated category.

59. A database according to claim 55, wherein the database structure provides said plurality of subject matter categories as a tree structure consisting of a plurality of main subject matter areas each divided into two or more subsidiary subject matter areas.

60. A database according to claim 55, wherein the database structure provides said plurality of subject matter categories such that each category is defined by a subject matter area and a species or genus.

61. A database according to claim 60, wherein the database provides said plurality of subject matter categories such that the species or geni are people, places, organisations, products and technology.

62. A database according to claim 60, wherein the database structure provides said plurality of subject matter categories such that the species or genus are the same for each subject matter area.

63. A database according to claim 59, wherein the database provides categories in each of the following subject matter areas: the universe, the earth, the environment, natural history, humanity, recreation, society, the mind and human history.

64. A database according to claim 59, wherein the database structure is such that, for a given meaning, a term is associated with only one category and different meanings of the same term are associated with different categories.

SUB A11

65. Apparatus for classifying electronic documents, comprising:

storage means storing a classification scheme having

a plurality of collocations each collocation being associated with a respective different subject matter area and containing a set of terms which exemplify that subject matter area;

means for comparing terms used in a document to be classified with the terms in said collocations;

means for allocating the document being classified to the one of said collocations which said comparing means identifies as having the most number of terms in common with the document being classified;

means for associating with the document being classified a code representing the subject matter area of the allocation collocation; and

means for storing the document together with the associated code.

66. Apparatus for filtering electronically stored documents forming the results of a search carried out by a search engine on the basis of a keyword supplied to the search engine by a user, comprising:

means storing a classification scheme divided into a number of collocations each associated with a specific different one of a number of different subject matter areas, each collocation containing a set of terms which exemplify the associated subject matter area;

means storing a vocabulary or dictionary of words with each word in the vocabulary being associated with one or more of said collocations, a description of the subject area of each associated collocation and a respective different definition of the word for each associated collocation;

means for determining from the vocabulary storing means each collocation with which the keyword is associated;

a user interface for providing the user with the subject area descriptions of each collocation with which the keyword is associated and for requesting the user to select one of said collocations; and

means responsive to the selection of a collocation by the user for comparing the terms contained in the selected collocation with terms used in each of the documents identified by the search engine and for providing the user with only those of said documents having more than a predetermined number of terms in common with the selected collocation.

67. A data carrier carrying a first set of data divided into a number of collocations each associated with a specific different one of a number of different subject matter areas with each collocation containing a set of

terms which exemplify the associated subject matter area, and a second set of data comprising a vocabulary or dictionary of terms with each entry in the vocabulary being associated with a respective different code associating it with a specific one of said collocations for each different context or meaning of the entry.

68. Apparatus for storing data on a computer-readable storage medium, comprising:

means for storing items of data;

means for associating each item of data with one of a number of subject matter areas such that each item of data belongs to at least one subject matter area;

means for associating each item of data with one of a number of different species areas or genera so that each item of data is associated with only one genus; and

means for directly or indirectly writing each item of data together with information identifying the associated subject matter area and genus onto a computer readable storage medium.

69. Apparatus for processing computer usable data, comprising:

means for storing items of data;

means for associating each item of data with at

least one of a number of different subject matter areas;

means for associating each item of data with only one of a number of species areas or genera; and

means for generating a signal carrying each item of data together with information identifying the associated subject matter area and genus.

SUB B14

70. A signal carrying processor implementable instructions for causing apparatus to become configured to form apparatus in accordance with claim 1.

71. A signal carrying processor implementable instructions for causing apparatus to become configured to form apparatus in accordance with claim 2.

72. A signal carrying processor implementable instructions for causing apparatus to become configured to form apparatus in accordance with claim 3.

SUB B15

73. A signal carrying a database in accordance with claim 53 or a plurality of collocations for use with the database.

74. A storage medium carrying a database in accordance with claim 53 or a plurality of collocations for use with the database.

75. A processor readable medium storing processor readable instructions for causing a processor to:

access a database having a database structure

providing a plurality of different subject matter categories, the database containing a classified vocabulary consisting of terms in all of the different subject matter categories with each term being classified in accordance with the subject matter category structure of the database and means for receiving in computer-readable form a text document to be classified;

compare terms appearing in the text document with the terms in the classified vocabulary;

determine from the comparison the category for the document; and

supply a signal carrying data representing the text document and data associating the text document with the determined category.

76. A processor readable medium storing processor readable instructions for causing a processor to:

access a database having a database structure providing a plurality of different subject matter categories, the database containing a classified vocabulary consisting of terms in all of the different subject matter categories, the database containing a classified vocabulary consisting of terms in all of the different subject matter categories with each term being classified in accordance with the subject matter category structure of the database;

receive a text document to be spell-checked;

compare terms appearing in the text document with the terms in the classified vocabulary;

determine from the comparison the category for the document;

identify any term in the document not present in the classified vocabulary; and

advise a user of the determined term or terms.

77. A processor readable medium storing processor readable instructions for causing a processor to:

access a database having a database structure providing a plurality of different subject matter categories, the database containing a classified vocabulary consisting of terms in all of the different subject matter categories with each term being classified in accordance with the subject matter category structure of the database;

receive documents forming the results of the subject matter search;

compare the keyword used to carry out the search with the classified vocabulary to determine each category which the keyword is associated;

advise a user of the different categories with one of the subject matter categories and each collocation consisting of a plurality of terms exemplifying the associated category.

78. A processor readable medium storing processor readable instructions for causing a processor to:

store terms;

associate each term with one of a number of different subject matter categories;

associate all terms falling within the same category with a common code identifying a collocation of terms exemplifying that category so that terms in different

categories are associated with different codes identifying different collocations with each collocation being specific to the associated category; and

supply as a database each term together with the associated code.

ADD B17